Instantaneous Throughput Prediction in Cellular Networks: Which Information Is Needed?

Alassane Samba*, Yann Busnel†, Alberto Blanc‡, Philippe Dooze* and Gwendal Simon‡
*Orange Labs, Lannion, France
†Crest (ENSAI) / Inria Rennes, France
‡IMT Atlantique / IRISA Rennes, France

***Abstract*—Downlink data rates can vary significantly in cellular networks, with a potentially non-negligible effect on the user experience. Content providers address this problem by using different representations (*e.g.*, picture resolution, video resolution and rate) of the same content and switch among these based on measurements collected during the connection. If it were possible to know the achievable data rate before the connection establishment, content providers could choose the most appropriate representation from the very beginning. We have conducted a measurement campaign involving 60 users connected to a production network in France, to determine whether it is possible to predict the achievable data rate using measurements collected, before establishing the connection to the content provider, on the operator's network and on the mobile node. We show that it is indeed possible to exploit these measurements to predict, with a reasonable accuracy, the achievable data rate.**

I. Introduction

In cellular networks, Quality of Service (QoS), in particular throughput, is especially sensitive to the context of use. To deal with changing QoS, content providers implement *adaptive* delivery strategies, where the quality and the characteristics of the delivered content are adjusted to match the achievable QoS of each user. These adaptive strategies are reactive: the characteristics of the content delivered at time t are based on measurements collected between the beginning of the connection and t.

Yet, content providers take some key decisions at the beginning of the delivery. For instance, most web services have several style sheets for their web pages, with a variable number of elements and information. The decision of which style sheet to deliver should be taken at the very beginning of the connection, even though no past throughput observation is available. Another example is adaptive video streaming. The video is divided into several chunks, and each chunk encoded at several bit-rate corresponding to different quality levels. Throughout the delivery, the client selects the highest quality representation with respect to an estimation of the available throughput based on the most recent history. At the very beginning, though, no such history is available. The delivery often starts with a medium or low quality representation to be on the safe side [1].

In both cases, content providers could avoid this guess work if they could get a reasonably accurate estimate of the achievable data rate with a given client. This estimate does not need to be extremely precise. Getting the order of magnitude can already be enough in most cases.

Several authors have proposed methods to predict the achievable data rate for a connection in a cellular network based on measurements collected during the connection itself. By their very nature, such methods cannot be used at the beginning of the connection. These methods use measurements collected over a certain time period, from a few milliseconds to a few minutes to make predictions over similar time scales. Some recent proposals have addressed throughput predictions based on the information about the radio link status [2]. This information is available at the mobile phone of the end-user; it allows thus instantaneous throughput prediction. However, the authors study accurate short-term prediction (a few hundreds of milliseconds at most), while the needs of content providers are also for rough throughput estimation at the scale of a few seconds (the length of a video chunk typically ranges from 2 to 10 seconds). Other proposals rely on instant measurements at the physical layer or on traffic monitoring at the cell level to infer the bandwidth of a user but they cannot accurately predict the achievable throughput for a connection, as this value can depend on a combination of all these factors.

In this paper we study instantaneous (*i.e.* history-less) prediction of the achievable throughput of a connection over a period of a few seconds. We are interested in identifying the set of information that enable relatively accurate predictions. Radio link information can be collected on the User Equipments (UEs) , *e.g.*, Received Signal Strength Indicator (RSSI), Reference Signal Received Quality (RSRQ), and Signal to Interference and Noise Ratio (SINR). Context information can also be collected at UE: location by Global Positioning System (GPS) coordinates, speed, terminal category and frequency band used. Finally, the network operator can offer information about the cellular network performance, including the average cell throughput, the average number of users, the connection success rate and the Block Error Ratio (BLER). One of the questions we address is whether this latter set of information brings significant improvement to the prediction.

We present and analyse the results of a measurement campaign involving a total of 5,700 connections over 350 different cells from a production network. We use supervised machine learning techniques to analyze the contribution of different measurements. Based on this analysis, we show that it is indeed possible to use instantaneous measurements collected in the cellular network and on the mobile to predict throughput, with a reasonable accuracy. We show that combining physical

978-3-901882-89-0 @2017 IFIP

layer measurements from the mobile with measurements from the cellular network enables a much better prediction.

II. Related works

Over the years, different methods have been proposed to estimate and predict the available bandwidth in a computer network (see, for example, the surveys by Prasad et al. [3] and Chaudhari and Biradar [4] and references wherein). These methods exploit timing or other characteristics of the packets belonging to a connection in order to estimate the available capacity. In other words, they can work only after the connection has been established. Instead, we are interested in predicting the available capacity *before* a connection is established, therefore we cannot use these solutions.

Our work is more closely related to studies that have shown that it is possible to predict the data-rate of a cellular connection by using measurements collected at the physical layer [2, 5–7]. Along similar lines, some authors have incorporated such data-rate predictions into adaptation algorithms for video transmission [8–10]. The key element shared by these papers is that, in cellular networks, UEs and base stations periodically exchange radio channel measurements, which are used by the base station to make scheduling decisions. To a varying degree, these papers propose to propagate this information to other layers and/or entities. For instance, CQIC [2] presents a new transport layer protocol based on a cross-layer design. While such an approach is possible, it calls for major changes, not only in the mobile nodes and in the cellular network but also in the Internet at large.

Our approach is not as radical: we collect data that are already available today in production networks and terminals. One key difference is that we propose to collect data from both user terminals and base stations. These measurements can be collected and combined by an ad-hoc element in the cellular network, similarly to what proposed by the EONA framework [11] or the DASH-Aware Network Element (DANE) element in the recent Server and Network Assisted DASH (SAND) standard.

III. Input Dataset

A. Measurement Campaign

We collected the data used in this paper thanks to 60 volunteers who have installed a dedicated application [12] on their UE and used it for two weeks, in February 2016. Throughout the day, as long as the terminal is turned on, the application periodically downloads a file from a remote server, using a production cellular network. The size of the file is 32 Mbits. The file download was done only when the cellular network used Long Term Evolution (LTE) technology. The server is in a well-provisioned data-center (in other words, the server cannot be the bottleneck of the connection). Figure 1 shows the different elements involved.

Every time a UE downloads the file, it generates an *entry*. This entry contains a timestamp and the time it took to download the full file, which we convert into the achieved throughput. It also contains several measurements logged from the UE Operating System (OS) and from the Radio Access Network (RAN) management system. We detail some of them in Section III-B below.

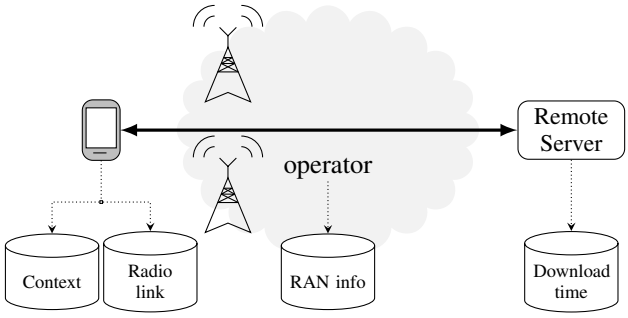

Fig. 1. Overall architecture of our data collection campaign

Finally, we filtered out the entries so that the entries having one or multiple missing values and the entries that used multiple base stations (*i.e.*, handover) are rejected. We obtained the entries of 31 volunteers, corresponding to 5,700 downloads on 350 different cells.

B. Dataset Description

We detail here the variables grouped into four families that we use as inputs of the prediction algorithm.

UE Categories and Cell Frequency Band LTE2600 and LTE800 base stations respectively correspond to cells using 2.6 GHz and 800 MHz frequency bands. LTE defines UE categories, which determine their performance specifications and enable base stations to be aware of their expected performance level. Only categories 3 and 4 have been used.

Physical Layer (Radio) On the UE we collect (i) *Reference Signal Received Power (RSRP)* the reference signal power across the channel bandwidth; and (ii) *Reference Signal Received Quality (RSRQ)* the ratio between Reference Signal Received Power (RSRP) and RSSI multiplied by the number of resources blocks allocated to the UE.

Context Information Intuitively, the awareness of the context in which the download operation occurs can help to predict QoS. Context awareness has been used in various other applications [13, 14]. In this paper, we consider the following indicators: (i) *Indoor/Outdoor*, an heuristic based on the number of visible GPS satellites from the UE reports whether the UE is indoor or not; (ii) *Distance to cell* based on the GPS coordinates provided by the application and a network topology database; and (iii) *Speed* estimated thanks to the GPS and the accelerometer.

RAN Measurements Operators use Network Manangement System (NMS) to monitor their networks by collecting raw counters of network events, typically aggregated over a period of a fifteen minutes. We have studied tens of KPIs, intuitively linked to the throughput, and we have concluded that the most relevant metrics are: average cell throughput, average number of users on the cell, BLER of the cell and Radio Resource

Control (RRC) setup success rate.

IV. Prediction Mechanism

In this part, we study which data are correlated with throughput and whether combining several data in inputs improves the throughput prediction.

A. Methodology

We used a Random Forest algorithm [15] for the learning technique. We used a K-fold method [16] with $K = 10$ for validation. It consists in dividing the set of entries into 10 randomly chosen subsets. Then we use 9 subsets to learn the best parameter settings for the predictor and the remaining subset is used as a test set. This well-known methodology enables to check if a model can accurately predict the throughput for a new entry.

The predictors are built as follows. First, we always include the UE category and the cell frequency band. Then, we consider the families of available data as described in Section III-B: Context, Radio link, and RAN. We look at all the configurations of availability for each family. Let i be an entry of our measurement campaign. Let y_i be the actual throughput of the download operation related to i. For each predictor, the algorithm predicts the throughput $\hat{y_i}$ and we compare the predicted throughput $\hat{y_i}$ to the actual achieved throughput y_i.

B. Performance Metrics

We have selected the two metrics, presented in Table I, among those commonly used to evaluate the results of prediction algorithms:

- The coefficient of determination, R^2, represents the percentage of the variance of the throughput explained by the predictor. It is calculated as follows: $R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}$, where $\bar{y}$ is the mean throughput.
- The median absolute error ratio, $\bar{E}_i$, is the error ratio that half of the predictions reach. The error ratio is measured by the absolute value of the difference between the predicted and the actual throughput, divided by the actual throughput.

To complete these results, we depict in Figure 2 the Empirical Cumulative Distribution Function (ECDF) of the prediction error ratio. We focus on the three main predictors: Radio-only, RAN-only, and Radio and RAN.

C. Results

Regarding the accuracy of the prediction, the results that we obtain (especially a cross validated coefficient of determination at 0.85 and a median error ratio at 0.1) are equivalent to much more sophisticated non-instantaneous techniques [17]. Our study thus reveals that instantaneous prediction based on data that are already available at the device and at the operator enables a suficient accurate prediction to allow content providers to select a *class of service* for each end-users.

The analysis of the best predictor should balance the accuracy (the higher the better) and the number of input data (the fewer the easier to implement). The first line of Table I shows that the cell band and the UE category do not enable an accurate prediction with our supervised learning technique. The context information allows an improvement, but the two main families of collected data that lead to a more accurate prediction are RAN and Radio link (the coefficient of determination is 0.70 and 0.71 respectively). Second, both RAN and Radio are complementary input data since the combination of both increases the R^2 to 0.85 and limits the median error ratio to 0.1.

TABLE I
Comparison of different predictors

Predictor	# variables	# entries	R^2	$\bar{E}_i$
x = UE cat. + Cell band	2	5757	0.39	0.28
x + Radio	4	4677	0.70	0.19
x + RAN	6	2842	0.71	0.17
x + Context	5	3827	0.65	0.20
x + Radio + RAN	8	2626	0.85	0.11
x + Radio + Context	7	3193	0.81	0.13
x + RAN + Context	9	1871	0.74	0.15
x + Radio + RAN + Context	11	1813	0.84	0.10

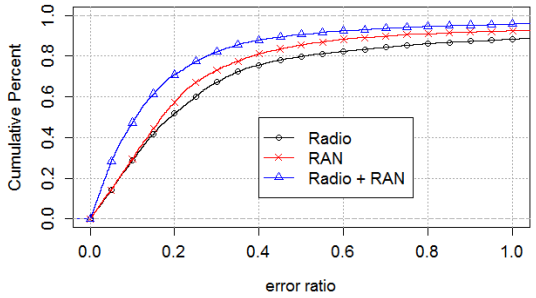

Fig. 2. Cumulative distribution function of error rate

V. Discussion

The achievable throughput of a connection over a cellular network depends on the performance of all the components involved in the transmission: the mobile device, the radio link, the cell capacity, the core network, and even the server of the content provider. A commonly accepted claim is that the network operator and the content provider over-provision the core network and Content Delivery Network (CDN) respectively so that the bottleneck is located in the *last-mile*. If this is indeed the case, predicting the data rate in the last-mile is equivalent to predicting the end-to-end data rate.

Some researchers have also studied the case where the core network [18] or the CDN [19] are under-provisioned, in which case the bottleneck is not in the last-mile. In this case, predicting the data rate on the wireless link is not enough to predict the end-to-end rate but such a prediction can still be exploited by combining it with information related to the status of the core network and the CDN. Typically, since content providers now use several CDNs [20] to deliver content, specific CDN monitoring solutions emerges. It would therefore be possible

to integrate the results of these monitoring solutions as inputs of our throughput prediction algorithm.

Another source of improvement for our algorithm is to use other types of information related to the mobile phone of the user. At the physical layer, the reception sensitivity and the transmission power are two variables that can impact the QoS.

Finally, a limitation of our study is the relatively poor availability of radio link information. The measurements about radio link are accessed by the mobile device OS through specific Application Programming Interfaces (APIs). Unfortunately, OS developers increasingly restrict these OS APIs. The Minimization of Drive Test (MDT) standard [21] can fix this problem by allowing the network operator to access the radio link information for each subscriber.

VI. Conclusion

Predicting a transmission throughput through cellular network using information available before the connection is a challenge. Our results confirm the correlation between throughput and both physical layer and access network data. We highlight how complementary are these inputs, which call for a better coordination between phone manufacturers, network operators and content providers. In our study, we collected information about the user context, cellular link quality, and access network performance data. These data are available before the connection at the condition that network operators and content providers share information. With a supervised learning technique, we have shown that it is possible to get an accurate prediction, which has the potential to help content providers to set their adaptive technique at the very beginning of the delivery. Our future work include the design of the collaboration between operators and content providers.

Acknowledgment

This work is supported by the EU project CogNet, 671625 (H2020-ICT-2014-2, Research and Innovation action).

References

[1] R. K. P. Mok, W. Li, and R. K. C. Chang, "Irate: Initial video bitrate selection system for HTTP streaming," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 6, pp. 1914–1928, 2016.

[2] F. Lu, H. Du, A. Jain, G. M. Voelker, A. C. Snoeren, and A. Terzis, "CQIC: Revisiting cross-layer congestion control for cellular networks," in *ACM HotMobile Workshop*, 2015.

[3] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy, "Bandwidth estimation: metrics, measurement techniques, and tools," *IEEE Network*, vol. 17, no. 6, pp. 27–35, Nov. 2003.

[4] S. S. Chaudhari and R. C. Biradar, "Survey of Bandwidth Estimation Techniques in Communication Networks," *Wireless Personal Communications*, vol. 83, no. 2, pp. 1425–1476, Mar. 2015.

[5] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 355–367, 2016.

[6] J. Yao, S. S. Kanhere, and M. Hassan, "An empirical study of bandwidth predictability in mobile computing," in *ACM WINTECH workshop*, 2008.

[7] Q. Xu, S. Mehrotra, Z. Mao, and J. Li, "PROTEUS: Network Performance Forecast for Real-time, Interactive Mobile Applications," in *ACM MobiSys*, 2013.

[8] X. K. Zou, J. Erman, V. Gopalakrishnan, E. Halepovic, R. Jana, X. Jin, J. Rexford, and R. K. Sinha, "Can accurate predictions improve video streaming in cellular networks?" in *ACM HotMobile*, 2015.

[9] X. Xie, X. Zhang, S. Kumar, and L. E. Li, "pistream: Physical layer informed adaptive video streaming over lte," in *ACM MobiCom*, 2015.

[10] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *ACM MobiCom*, 2013.

[11] J. Jiang, X. Liu, V. Sekar, I. Stoica, and H. Zhang, "EONA: Experience-Oriented Network Architecture," in *ACM HotNet*, 2014.

[12] "EQual One - http://www.v3d.fr/solution/equal-one/," Vision 360 Degres (V3D), Feb. 2016.

[13] X. Hu, X. Li, E. C. H. Ngai, V. C. M. Leung, and P. Kruchten, "Multidimensional context-aware social network architecture for mobile crowdsensing," *IEEE Communications Mag.*, vol. 52, no. 6, pp. 78–87, 2014.

[14] S. Hahn, D. Gotz, S. Lohmuller, L. Schmelz, A. Eisenblätter, and T. Kürner, "Classification of cells based on mobile network context information for the management of SON systems," in *IEEE VTC Spring*, 2015.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 1995.

[17] M. Mirza, J. Sommers, P. Barford, and X. Zhu, "A machine learning approach to TCP throughput prediction," in *ACM Sigmetrics conference*, 2007.

[18] S. B. H. Said, M. R. Sama, K. Guillouard, L. Suciu, G. Simon, X. Lagrange, and J.-M. Bonnin, "New Control Plane in 3GPP LTE/EPC Architecture for On-Demand Connectivity Service," in *IEEE CloudNet*, 2013.

[19] J. Liu, G. Simon, C. Rosenberg, and G. Texier, "Optimal delivery of rate-adaptive streams in underprovisioned networks," *IEEE Journal on Selected Areas in Communications*, 2014.

[20] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang, "Unreeling netflix: Understanding and improving multi-cdn movie delivery," in *IEEE INFOCOM*, 2012.

[21] ETSI, "Trace data definition and management," *3GPP TS 32.423 version 13.0.0 Release 13*, 2016.